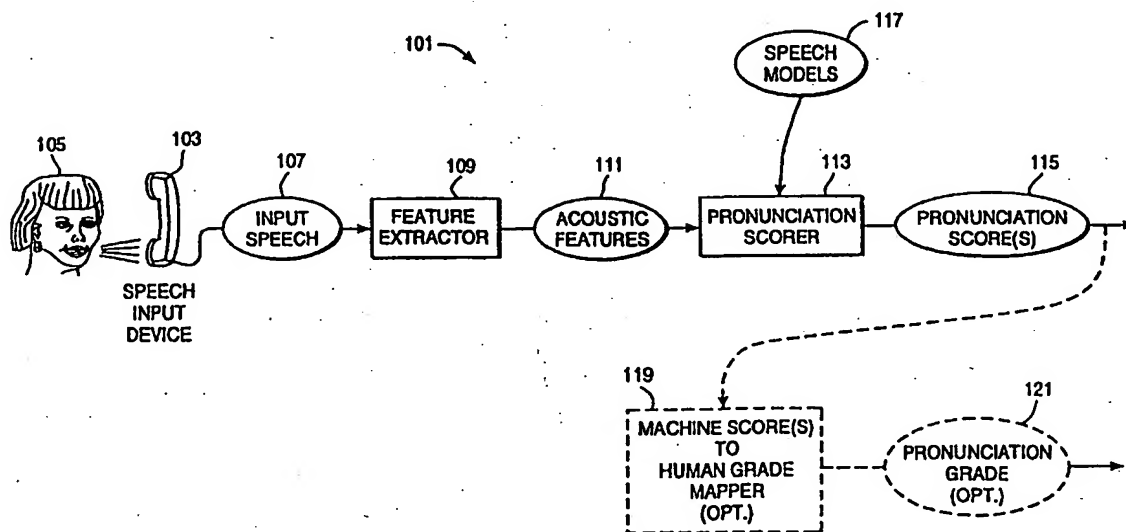


INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : G10L 5/04, 9/00, G09B 5/00	A1	(11) International Publication Number: WO 98/14934 (43) International Publication Date: 9 April 1998 (09.04.98)
(21) International Application Number: PCT/US97/17888 (22) International Filing Date: 1 October 1997 (01.10.97) (30) Priority Data: 60/027,638 2 October 1996 (02.10.96) US Not furnished 23 September 1997 (23.09.97) US (71) Applicant: SRI INTERNATIONAL [US/US]; 333 Ravenswood Avenue, Menlo Park, CA 94025 (US). (72) Inventors: NEUMEYER, Leonardo; 3428 South Court, Palo Alto, CA 94306 (US). FRANCO, Horacio; 197 Ravenswood Avenue, Atherton, CA 94027 (US). WEINTRAUB, Mitchel; 36360 Coronado Drive, Fremont, CA 94536 (US). PRICE, Patti; 420 Shirley Way, Menlo Park, CA 94025 (US). DIGALAKIS, Vassilios; Kalamaki 11, N. Kydonia, Chania 73100 (GR). (74) Agents: ALLEN, Kenneth, R. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).		(81) Designated States: JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD AND SYSTEM FOR AUTOMATIC TEXT-INDEPENDENT GRADING OF PRONUNCIATION FOR LANGUAGE INSTRUCTION



(57) Abstract

Acoustic features (109, 111) are extracted from input speech (107) and are compared (113) against pre-stored models (117). The result is used to make a judgement of the user's pronunciation (115).

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

5 **METHOD AND SYSTEM FOR AUTOMATIC TEXT-INDEPENDENT
GRADING OF PRONUNCIATION FOR LANGUAGE INSTRUCTION**

10 **STATEMENT OF RELATED APPLICATIONS**

10 This patent application claims priority from U.S.
Provisional Application No. 60/027,638, filed 10/2/96. The
content of the provisional application is incorporated herein
by reference.

15 **COPYRIGHT NOTICE**

15 A portion of the disclosure of this patent document
contains material which is subject to copyright protection.
The copyright owner has no objection to the facsimile
20 reproduction by anyone of the patent document or the patent
disclosure as it appears in the Patent and Trademark Office
patent file or records, but otherwise reserves all copyright
rights whatsoever.

25 **BACKGROUND OF THE INVENTION**

25 The present invention relates to automatic
evaluation of speech pronunciation quality. One application
is in computer-aided language instruction and assessment.

30 Techniques related to embodiments of the present
invention are discussed in co-assigned U.S. Application No.
08/375,908, entitled METHOD AND APPARATUS FOR SPEECH
RECOGNITION ADAPTED TO AN INDIVIDUAL SPEAKER; U.S. Application
No. 08/276,742, entitled METHOD AND APPARATUS FOR SPEECH
35 RECOGNITION USING OPTIMIZED PARTIAL MIXTURE TYING; U.S. Patent
No. 5,634,086, entitled METHOD AND APPARATUS FOR VOICE-
INTERACTIVE LANGUAGE INSTRUCTION; and U.S. Patent No.
5,581,655, entitled METHOD FOR RECOGNIZING SPEECH USING
LINGUISTICALLY-MOTIVATED HIDDEN MARKOV MODELS; which
40 applications and patents are incorporated herein by reference.

Relevant speech recognition techniques using Hidden Markov Models are also described in V. Digalakis and H. Murveit, "GENONES: Generalized Mixture-Tying in Continuous Hidden-Markov-Model-Based Speech Recognizers," IEEE Transactions on Speech and Audio Processing, Vol. 4, July, 1996, which is incorporated herein by reference.

Computer-aided language instruction systems exist that exercise the listening and reading comprehension skills of language students. While such systems have utility, it would be desirable to add capabilities to computer-based language instruction systems that allow students' language production skills also to be exercised. In particular, it would be desirable for a computer-based language instruction system to be able to evaluate the quality of the students' pronunciation.

A prior-art approach to automatic pronunciation evaluation is discussed in previous work owned by the assignee of the present invention. See Bernstein et al., "Automatic Evaluation and Training in English Pronunciation", Internat. Conf. on Spoken Language Processing, 1990, Kobe, Japan. This prior-art approach is limited to evaluating speech utterances from students who are reading a pre-selected set of scripts for which training data had been collected from native speakers. This prior-art approach is referred to as text-dependent evaluation because it relies on statistics related to specific words, phrases, or sentences.

The above-referenced prior-art approach is severely limited in usefulness because it does not permit evaluation of utterances which were not specifically included in the training data used to train the evaluation system, so that retraining of the evaluation system is required whenever a new script needs to be added for which pronunciation evaluation is desired.

What is needed are methods and systems for automatic assessment of pronunciation quality capable of grading even arbitrary utterances--i.e., utterances made up of word sequences for which there may be no training data or

incomplete training data. This type of needed pronunciation grading is termed text-independent grading.

5 The prior-art approach is further limited in that it can generate only certain types of evaluation scores, such as a spectral likelihood score. While the prior-art approach achieves a rudimentary level of performance using its evaluation scores, the level of performance is rather limited, as compared to that achieved by human listeners. Therefore, what is also needed are methods and systems for automatic
10 assessment of pronunciation quality that include more powerful evaluation scores capable of producing improved performance.

GLOSSARY

15 In this art, the same terms are often used in different contexts with very different meanings. For purposes of clarity, in this specification, the following definitions will apply unless the context demands otherwise:

Grade: An assessment of the pronunciation quality
20 of a speaker or a speech utterance on a grade scale such as used by human expert listeners. A grade may be human- or machine-generated.

Score: A value generated by a machine according to a scoring function or algorithm as applied to a speech
25 utterance.

A Frame of Acoustic Features: A characterization of speech sounds within a short time-frame produced by a feature extractor for subsequent processing and analysis. For example, a feature extractor that computes acoustic features
30 every 10 ms within a shifting 20 ms window is said to produce a "frame of acoustic features" every 10 ms. In general, a frame of acoustic features is a vector.

Acoustic Segments: Time-segments of speech whose boundaries (or durations) are determined by a speech segmenter
35 based on acoustic properties of the speech. In an embodiment of the invention, each acoustic segment produced by the speech segmenter is a "phone."

Phone: A basic speech sound unit within a given language. In general, all speech utterances for a given language may be represented by phones from a set of distinct phone types for the language, the number of distinct phone types being on the order of 40.

Acoustic Units: Time-segments of speech whose durations are used to generate a score that is indicative of pronunciation quality. In an embodiment of the invention, acoustic units are simply the acoustic segments produced by the speech segmenter. In another embodiment, acoustic units are "syllables" whose durations are determined based on the boundaries (or durations) of the acoustic segments produced by the speech segmenter.

SUMMARY OF THE INVENTION

According to the invention, methods and systems are provided for assessing pronunciation quality of an arbitrary speech utterance based on one or more metrics on the utterance, including acoustic unit duration and a posterior-probability-based evaluation.

A specific embodiment of the invention is a method for assessing pronunciation of a student speech sample using a computerized acoustic segmentation system, wherein the method includes: accepting the student speech sample which includes a sequence of words spoken by a student speaker; operating the computerized acoustic segmentation system to define acoustic units within the student speech sample based on speech acoustic models within the segmentation system, the speech acoustic models being established using training speech data from at least one speaker, the training speech data not necessarily including the sequence of spoken words; measuring duration of the sample acoustic units; and comparing the sample acoustic unit durations to a model of exemplary acoustic unit duration to compute a duration score indicative of similarity between the sample acoustic unit durations and exemplary acoustic unit durations.

According to a further specific embodiment, the duration score is further mapped to a grade, and the grade is presented to the student speaker.

According to a further specific embodiment, the spoken sequence of words is unknown, and a computerized speech recognition system is operated to determine the spoken sequence of words.

A further specific embodiment of the invention is a method for grading the pronunciation of a student speech sample, the method including: accepting the student speech sample which includes a sequence of words spoken by a student speaker; operating a set of trained speech models to compute at least one posterior probability from the speech sample, each of the posterior probabilities being a probability that a particular portion of the student speech sample corresponds to a particular known model given the particular portion of the speech sample; and computing an evaluation score, herein referred to as the posterior-based evaluation score, for the student sample of pronunciation quality from the posterior probabilities.

According to a further specific embodiment, the posterior-based score is further mapped to a grade as would be assigned by human grader, and the grade is presented to the student speaker.

A still further specific embodiment of the invention is a system for assessing pronunciation of a student speech sample, the student speech sample including a sequence of words spoken by a student speaker, the system including: trained speech acoustic models of exemplary speech; and an acoustic scorer configured to compute at least one posterior probability from the speech sample using the trained speech models, the acoustic scorer also configured to compute an evaluation score of pronunciation quality for the student sample from the posterior probabilities, each of the posterior probabilities being a probability that a particular portion of the student speech sample corresponds to a particular known model given the particular portion of the speech sample.

A still further specific embodiment of the invention is a system for pronunciation training in a client/server environment wherein there exists a client process for presenting prompts to a student and for accepting student speech elicited by the prompts, the system including: a server process for sending control information to the client process to specify a prompt to be presented to the student and for receiving a speech sample derived from the student speech elicited by the presented prompt; and a pronunciation evaluator invocable by the server process for analyzing the student speech sample.

A further understanding of the nature and advantages of the present invention may be realized by reference to the remaining portions of the specification and the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a system for evaluating pronunciation quality.

FIG. 2 is a block diagram of a pronunciation scorer of FIG. 1 that produces a pronunciation score based on duration of acoustic units according to an embodiment of the present invention.

FIG. 3 is a block diagram showing a speech segmenter of FIG. 2 that is a hidden Markov model (HMM) speech recognizer according to an embodiment of the present invention.

FIG. 4 is a diagram illustrating a portion of a maximum likelihood path for sample input speech.

FIG. 5 is a block diagram of a system for computing an acoustic score based directly on the acoustic features themselves according to embodiments of the present invention.

FIG. 6 is a block diagram of a system that combines different pronunciation scores according to an embodiment of the invention.

FIG. 7 is a block diagram of a system for creating FIG. 6's mapping function between one or more types of machine

scores into a pronunciation grade as would be produced by a human grader.

FIG. 8 is a block diagram of a distributed language instruction system that evaluates pronunciation quality.

DESCRIPTION OF SPECIFIC EMBODIMENTS

I. AUTOMATIC PRONUNCIATION EVALUATION

FIG. 1 is a block diagram of a system 101 for evaluating pronunciation quality according to embodiments of the present invention. In FIG. 1, a speech input device 103 converts a sequence of spoken words from a speaker 105 into machine-readable input speech 107. A feature extractor 109 divides the input speech 107 into time-frames and computes, for each time-frame, acoustic features that capture distinguishing characteristics of speech sounds within the time-frame. In this manner, the feature extractor 109 produces a sequence of acoustic feature frames 111. The input speech 107 and the sequence of acoustic feature frames 111 are both representations of the speaker 105's speech and may therefore each be referred to as a "student speech sample."

A pronunciation scorer 113 computes from the acoustic features 111 at least one pronunciation score 115 that is indicative of pronunciation quality of the input speech 107. In computing the pronunciation scores 115, the pronunciation scorer 113 relies upon speech models 117 which characterize various aspects of desirable, i.e. exemplary, speech pronunciation. The speech models 117 are established using training speech from exemplary speakers.

In some embodiments of the present invention, an optional score-to-grade mapper 119 accepts the pronunciation scores 115 and maps them into a pronunciation grade 121 as would be given by an expert human grader.

During operation of the pronunciation evaluation system 101, the various data, including the input speech 107, the acoustic features 111, the pronunciation score(s) 115, and

the pronunciation grade 121 may be stored in storage devices for later use.

In embodiments of the present invention, the acoustic features 111 include features used in the speech recognition task, which are known in the art and are discussed for example in the references cited and referenced in the Background section. For example, in an embodiment of the present invention, the acoustic features 111 include 12th order mel-cepstra features computed every 10 ms within a shifting 20 ms window, and the features' approximate derivatives.

In an embodiment of the present invention, the speech input device 103 is a telephone, and the speech input 107 is conveyed across a telephone network to the feature extractor 109. This embodiment enables students to have their spoken pronunciation evaluated by the present invention so long as they have access to a telephone.

In an embodiment of the present invention, the speech input device 103 is a digitizing microphone system, such as a microphone connected to a remote, "client" computing system that contains hardware and software for audio digitization. The input speech 107 is conveyed in digitized form (e.g., as streaming audio or as a compressed audio file) across a digital network, for example, a local area network and/or the Internet, to the feature extractor 109 which exists on a local, "server" computing system. This embodiment enables students to have their spoken pronunciation evaluated by the present invention so long as they have access to a digitizing microphone system connected to the digital network.

In an embodiment of the present invention, the speech input device 103 and the feature extractor 109 reside on at least one remote computing system and the acoustic features 111 are conveyed across a network, for example, the Internet, to the pronunciation scorer 113 which exists on a local computing system. This embodiment reduces the amount of data which need be conveyed across the network because acoustic features 111 typically are a more compact representation of speech than are the input speech 107 itself

in this embodiment. This embodiment also reduces the amount of computation required of the local computing system.

II. SCORING PRONUNCIATION USING ACOUSTIC UNIT DURATIONS

FIG. 2 is a block diagram of a pronunciation scorer 113 of FIG. 1 according to embodiments of the present invention that produce a pronunciation score 115 based on duration of acoustic units. In FIG. 2, a speech segmenter 203 accepts the sequence of acoustic features 111 and produces from them a time-segmentation 205 specifying acoustic segments. The acoustic segmentation 205 is a representation of acoustic segments from which their durations may be determined. In an embodiment, the acoustic segmentation 205 comprises a time-boundary of each acoustic segment plus each acoustic segment's duration. (Note that in general segment boundaries define durations, and a sequence of durations defines segment boundaries given a single boundary within the sequence. Therefore, a system component that is described as using boundaries can in general be alternatively but equivalent described as using durations, or durations plus a boundary.)

An acoustic unit duration extractor 207 accepts the acoustic segmentation 205. From the acoustic segmentation 205, the acoustic unit duration extractor 207 recovers or computes durations 209 of the acoustic units.

An acoustic unit duration scorer 211 accepts the acoustic unit durations 209 and compares them to a model 213 of exemplary acoustic unit duration which has been established using training speech from exemplary speakers. Based on this comparison, the acoustic unit duration scorer 211 computes an acoustic unit duration score 115 as the pronunciation score 115 of FIG. 1. The acoustic unit duration model 213 forms a part of the speech models 117 of FIG. 1. In embodiments of the invention, the acoustic unit duration model 213 may be a parametric model or a non-parametric model. In another embodiment of the invention, the acoustic unit duration model

213 simply contains example acoustic unit durations from exemplary speech.

It has been found that acoustic unit duration scores are particularly important indicators of pronunciation quality when the student speaker 105's speech is received through a channel that adds a large amount of noise or distortion, such as speech transmitted through a telephone connection.

In an embodiment of the present invention, the speech input device 103 (of FIG. 1), the feature extractor 109 (of FIG. 1), and the speech segmenter 203 all reside on one or more remote computing system(s) and only the acoustic segmentation 205 or only the acoustic unit durations 209 are conveyed across a network, for example, the Internet, to the acoustic unit duration scorer 211, which resides on a local computing machine. This embodiment drastically reduces the amount of data which need to be conveyed across the network and the amount of computation required of the local computing system, at the expense of requiring the remote computing system to perform more computations.

In embodiments of the present invention, the speech segmenter 203 segments the acoustic features 111 into acoustic segments which are phones. The speech segmenter 203 also identifies the type of each phone. The acoustic segmentation 205 includes segment information in the form of, for example, phone boundaries expressed as indices into the sequence of acoustic features 111 and phone type labels for each phone.

II.A. PHONE DURATION

Certain embodiments of the present invention compute duration scores 115 based on phone duration. The speech segmenter 203 segments the acoustic features 111 into acoustic segments which are phones. The acoustic unit duration extractor 207 defines acoustic units as, simply, the phones themselves. Therefore, the acoustic unit duration extractor 207 in these embodiments very simply extract the phone durations as the acoustic unit durations 209. In particular, in embodiments whose phone segmentation 205 expressly include

phone durations, the acoustic unit duration extractor 207 simply uses the existing phone durations as the acoustic unit durations 209. In embodiments whose phone segmentation 205 represents phone segmentation with only phone boundaries, the acoustic unit duration extractor 207 is an arithmetic subtractor that computes acoustic unit durations from the phone boundaries.

In certain phone-duration-scoring embodiments of the invention, the acoustic unit duration model 213 includes a separate probability distribution $P_d(d|q)$ of phone duration d in exemplary speech given the phone's type q . For example, a system configured to use, e.g., 45 types of phones that describe a given language would have 45 probability distributions, one for each phone type.

In a specific embodiment, each phone type's duration probability distribution is represented as a parametric distribution, such as a Gaussian distribution. The parameters of these distributions are estimated according to standard statistical estimation methods using the durations of each type of phone as found in training speech from exemplary speakers.

In other, preferred embodiments, each phone type's duration probability distribution is represented as a (nonparametric) probability mass function. These probability distributions are established by tabulating durations of each type of phone as found in training speech from exemplary speakers. Each probability mass function is smoothed, and a probability floor is introduced, in order to maintain robustness of the model, given that only finite quantities of training speech are available. Phone durations of the training speech are determined during training in the same manner as are phone durations 209 of input speech 107 determined during testing. Namely, the feature extractor 109, speech segmenter 203, and acoustic unit duration extractor 207 are used.

The acoustic unit duration scorer 211 in certain phone-duration-scoring embodiments computes a log-probability ρ_i of the duration d_i of each phone i :

$$\rho_i = \log P_d(d_i | q_i) \quad (1)$$

wherein q_i is the phone type of phone i .

The acoustic unit duration scorer 211 computes an acoustic unit duration score 115 ρ for an entire utterance as the average of the log-probability ρ_i of each phone i 's duration:

$$\rho = \frac{1}{N} \sum_{i=1}^N \rho_i \quad (2)$$

wherein the sum runs over a number N of phones in the utterance.

In a preferred embodiment, the acoustic unit duration model 213 includes probability distributions $P_{d'}(d' | q)$ of phone durations d' that are *speaker-normalized* phone durations. Accordingly, the acoustic unit duration scorer 211 computes the acoustic unit duration score 115 for an entire utterance as the average of the log-probability of each phone i 's speaker-normalized duration d'_i .

A speaker-normalized phone duration is the phone duration multiplied by the rate of speech for the speaker in question. Rate of speech (ROS) is the number of phones uttered by a speaker per second of speaking. The rate of speech of each exemplary speaker is calculated from the training speech. The rate of speech of the student speaker 105 is calculated from available data for the speaker, including the acoustic segmentation 205 itself.

The following equations summarize use of speaker-normalized phone durations in the preferred embodiment:

$$d'_i = d_i \text{ ROS} \quad (3)$$

$$\rho_i = \log P_{d'}(d'_i | q_i) \quad (4)$$

$$\rho = \frac{1}{N} \sum_{i=1}^N \rho_i \quad (2)$$

11.B. SYLLABIC DURATION

Certain embodiments of the present invention compute duration scores 115 based on the duration of "syllables". One explanation of why syllabic duration is a good indicator of pronunciation quality, even after normalization for rate of speech (as will be described), is that language learners tend to impose the rhythm of their native language on the language they are learning. For example, English tends to be *stress-timed* (stressed syllables tend to be lengthened and others shortened), while Spanish and French tend to be *syllable-timed*.

In these syllabic-duration-scoring embodiments, the acoustic unit duration extractor 207 determines durations of acoustic units that are "syllables" based on the durations of phones as specified by the speech segmenter 203. In particular, the acoustic unit duration extractor 207 determines syllabic durations as the duration between the centers of vowel phones within speech.

In a specific syllabic-duration-scoring embodiment, the acoustic unit duration model 213 includes a single probability distribution $P_{sd}(sd)$ of the syllabic duration sd of any syllable. This probability distribution is established by tabulating durations of all syllables found in training speech from exemplary speakers. Syllable durations of the training speech are determined during training in the same manner as are syllable durations 209 of input speech 107 determined during testing. Namely, the feature extractor 109, speech segmenter 203, and acoustic unit duration extractor 207 are used. The duration probability distribution is represented as a probability mass function. The probability mass function is smoothed, and a probability floor is introduced, in order to maintain robustness of the model, given that only finite quantities of training speech are available.

In a preferred embodiment, the syllabic duration sd_j for each syllable j is normalized during testing and training by multiplication with the speaker's rate of speech (ROS), as defined above, to obtain a speaker-normalized syllabic

duration sd'_j . The following equations summarize use of speaker-normalized syllabic durations in the preferred syllabic-duration-scoring embodiment:

$$sd'_j = sd_j ROS \quad (6)$$

$$\rho_j = \log P_{sd'}(sd'_j) \quad (7)$$

$$\rho = \frac{1}{M} \sum_{j=1}^M \rho_j \quad (8)$$

II.C. SYLLABIC DURATION USING SPECIFIC SYLLABLES

In other embodiments of the present invention, syllabic duration of specific syllables are used for scoring in a manner analogous to that described above for all syllables. In these embodiments, the acoustic unit duration extractor 207 recovers syllabic durations from the acoustic segmentation 205. The duration scorer compares these durations to a model 213 of syllabic duration in exemplary speech to compute a syllabic duration score 115.

The syllabic duration model 213 includes a probability distribution of duration for a subset of the syllables in the language. These syllables are the ones for which sufficient training speech data existed from which duration distributions could be estimated. The duration scorer compares syllables from the student speech sample with the syllable duration model 213 to derive syllabic duration pronunciation scores, based on those syllables of the student speech sample whose durations are modelled within the syllabic duration model 213.

II.D. WORD DURATION

In other embodiments of the present invention, word duration is used for scoring in a manner analogous to that described above for syllables. In these embodiments, the acoustic unit duration extractor 207 recovers word durations

from the acoustic segmentation 205. The duration scorer compares these durations to a model 213 of word duration in exemplary speech to compute a word duration score 115.

The word duration model 213 includes a probability distribution of duration for a subset of the words in the language. These words are the ones for which sufficient training speech data existed from which duration distributions could be estimated. The duration scorer compares words from the student speech sample with the word duration model 213 to derive word duration pronunciation scores, based on those words of the student speech sample whose durations are modelled within the word duration model 213.

III. AN HMM SPEECH RECOGNIZER FOR ACOUSTIC SEGMENTATION

FIG. 3 is a block diagram showing a speech segmenter 203 of FIG. 2 that is an HMM speech recognizer 203, according to a specific embodiment of the present invention. HMM speech recognizers are known in the art and are discussed for example in the references cited and incorporated in the Background section.

A Markov model (MM) is a network of states connected by directed transition branches. The HMM speech recognizer 203 uses a Markov model to model the production of speech sounds. The HMM recognizer 203 represents each type of phone in a language by a phone model made up of a handful of connected states. (The specific embodiment uses 3 states per phone model for most phone types.) The HMM recognizer 203 also provides additional, context-dependent phone models, including "tri-phone" models, that represent each phone type when it is preceded and/or followed by particular other phone types. The HMM recognizer 203 also includes a pause phone which models pauses that occur during speech between words. The phone models, including the context-dependent and pause phone models, form acoustic models 305 within the HMM recognizer 203.

Each state in a speech HMM has an associated probability distribution of the acoustic features which are

produced while in the state. (These output distributions are alternatively but equivalently described in the literature as being associated with the transition branches.) The output distributions may be Gaussian distributions, or weighted mixtures of Gaussian distributions, etc., as are described in the literature. The HMM recognizer 203 of the specific embodiment uses output distributions which are weighted tied mixtures of Gaussian distributions. Weighted tied mixtures are known in the art of speech recognition. A standard HMM speech recognizer which can be configured to implement the HMM recognizer 203 of the specific embodiment is the DECIPHER system from SRI International of Menlo Park, California.

Each transition branch in a Markov model has a transition probability indicating the probability of transiting from the branch's source state to its destination state. All transition probabilities out of any given state, including any self transition probabilities, sum to one.

The output and transition probability distributions for all states in a speech HMM are established from training speech data using standard HMM training algorithms and techniques, including the forward-backward (Baum-Welch) algorithm. A standard HMM-based speech recognizer on which such training can be performed is the DECIPHER system from SRI International of Menlo Park, California.

According to the present invention, the training speech are not required to include the sequence of spoken words found in the input speech 107. These training speech are not even required to include individual words from the sequence of spoken words found in the input speech 107.

A lexicon 307 is a catalog of words in a language and defines component phone types that make up each word. In some embodiments of the invention, the lexicon 307 also includes any assigned transition probabilities from phone type to phone type within each word. A grammar 309 describes allowed word-to-word transitions in a language. The grammar 309 of the specific embodiment is a "bi-gram" that specifies context-free word-to-word transition probabilities between every pair of words. The grammar 309 also allows an optional

pause phone between words to model possible pauses between words during speech. The grammar 309 allows the pause phone to be skipped. The grammar 309 implements a skip as a transition arc which does not correspond to any outputted acoustic features.

Grammars 309 and lexicons 307 together form a grammar network 310 that specifies allowable links between phones and, hence, allowable words and sentences. Grammars, lexicons, and grammar networks are known elements of HMM speech recognizers. The grammar network 310 and the phone acoustic models 305 form a part of the speech models 117 (of FIG. 1).

All phone models 305 plus the lexicon 307 and the grammar 309 may be considered to be a vast virtual network called "the HMMs" or "the recognition HMM." The HMM recognizer 203 models every spoken sentence as having been produced by traversing a path through the states within the HMMs. In general, a frame of acoustic features is produced at each time-step along this path. (However, some state transitions such as the "skip" transition take no time and produce no output.) The path identifies the sequence of states traversed. The path also identifies the duration of time spent in each state of the sequence, thereby defining the time-duration of each phone and each word of a sentence. Put in another way, the path describes an "alignment" of the sequence of frames 111 with a corresponding sequence of states of the HMMs.

In FIG. 3, the HMM speech recognizer 203 is operated not merely for its ordinary purpose of speech recognition, but also for time-segmenting speech into component phones. In FIG. 3, The HMM recognizer 203 accepts the acoustic features 111. The HMM recognizer 203 includes hidden Markov models (HMMs) specified by the phone acoustic models 305, the lexicon 307, and the grammar 309. An HMM search engine 311 within the HMM recognizer 203 computes a maximum likelihood path 313.

The maximum likelihood path is a path through the hidden Markov models with the maximum likelihood of generating the acoustic feature sequence 111 extracted from the speech of

the user. The maximum likelihood path 313 includes the sequence of states traversed 314 and the duration of time 315 spent in each state. The maximum likelihood path 313 defines an acoustic segmentation 205 of the acoustic features into a sequence of phones. The acoustic segmentation 205 of the specific embodiment is a subset of the path information 313, including time boundaries (and/or durations) and the phone-type labels of the sequence of phones. Using the duration information from the acoustic segmentation 205, the present invention evaluates pronunciation quality, as was described above in connection with Figs. 1 and 2.

The HMM search engine 311 computes the maximum likelihood path through its speech HMMs according to a standard pruning HMM search algorithm that uses the well-known Viterbi search method. This HMM search algorithm is described for example in the cited and incorporated art and elsewhere in the literature. The Viterbi algorithm is also discussed in numerous other references, such as G.D. Forney, Jr., "The Viterbi algorithm," Proc. IEEE, vol.61, pp.268-278, 1973.

In the specific embodiment, the sequence of spoken words from the speaker 105 may or may not be known in advance by the pronunciation evaluation system 101. If the sequence of spoken words is not known in advance, then the HMM recognizer 203 outputs, in addition to the acoustic segmentation 205, the recognized word sequence 317 for other use. For example, the recognized word sequence 317 may be used by an interactive language instruction system included in the specific embodiment. This language instruction system might determine the meaning of the recognized word sequence 317 and whether the recognized word sequence 317 is a correct and appropriate utterance in relation to a current lesson being conducted.

If the sequence of spoken words is known in advance, then the known word sequence 319 is fed to the HMM engine 311 to dramatically constrain the possible paths through the HMMs. This known word sequence 319 represents additional information that forms a part of the grammar network 310. The sequence of spoken words may be known in advance for example because a

language instruction system has requested that the speaker 105 read from a known script. Using the known word sequence 319 as an additional constraint can reduce recognition and segmentation errors and also reduce the amount of computation required by the HMM engine 311.

FIG. 4 is a diagram illustrating a portion of a maximum likelihood path 313 for sample input speech 107 in accordance with the invention. The input speech 107 is composed of its constituent words 403, which are in turn broken down into constituent phones 205, which are themselves broken down into constituent states 405. The constituent phones 205 include phone type labels 407 as well as information that specifies each phone's duration.

IV. SCORING PRONUNCIATION USING ACOUSTIC FEATURES

FIG. 5 is a block diagram of a system 113 for computing an acoustic score 115 based directly on the acoustic features 111 themselves, rather than on acoustic unit durations, according to embodiments of the present invention.

In FIG. 5, a speech segmenter 203 accepts the sequence of acoustic features 111 and produces from them a time-segmentation 205 specifying acoustic segments. An acoustic scorer 503 accepts the acoustic segmentation 205 and also the sequence of acoustic features 111. The acoustic scorer 503 uses the acoustic segmentation 205 to index into the sequence of acoustic features 111. In this way, the acoustic scorer 503 obtains acoustic feature frames which correspond to each acoustic segment.

The acoustic scorer 503 compares the acoustic feature frames of the acoustic segments to a model 505 of exemplary acoustic feature frames. The model 505 was established using training speech from exemplary speakers. Based on this comparison, the acoustic scorer 503 computes the acoustic score 115 as the pronunciation score 115 of FIG. 1. The acoustic model 505 forms a part of the speech models 117 of FIG. 1.

In certain acoustic-scoring embodiments of the invention, the speech segmenter 203 is an HMM recognizer 203 that produces an acoustic segmentation 205 of the sequence of acoustic features 111 into phones, as was described in connection with FIG. 3. The acoustic model 505 in certain of these embodiments includes separate models of acoustic feature frames for each phone type. In a preferred embodiment, these models are HMM models from the HMM recognizer 203 used for segmentation.

IV.A. PHONE LOG-POSTERIOR PROBABILITY SCORES

In a specific acoustic-scoring embodiment, each of the separate models corresponding to a phone type q is a context-independent probability density $p(y|q)$, wherein the variable y represents an acoustic feature frame. The acoustic scorer 503 computes, for each frame y_i within a phone i of phone type q_i , a frame-based posterior probability $P(q_i|y_i)$ of phone i 's type given the observed acoustic feature frame y_i :

$$P(q_i|y_i) = \frac{p(y_i|q_i) P(q_i)}{\sum_{\text{All } q} p(y_i|q) P(q)} \quad (9)$$

wherein $p(y_i|q_i)$ is the probability of the frame y_i according to the distribution corresponding to phone type q_i . The sum over q runs over all phone types. $P(q_i)$ represents the prior probability of the phone type q_i .

The acoustic scorer 503 of the specific embodiment computes a phone posterior score ρ_i for each phone i defined by the acoustic segmentation 205. Each phone i 's phone posterior score is an average of the logarithms of the frame-based posterior probabilities $P(q_i|y_i)$ of all frames within phone i . Each phone i 's phone posterior score ρ_i can be expressed as:

$$\rho_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log P(q_i|y_t) \quad (10)$$

wherein the sum runs over all d_i frames of phone i .

The acoustic scorer 503 of the specific embodiment computes the acoustic score 115 ρ for an entire utterance as the average of the phone posterior scores ρ_i of each phone i :

$$\rho = \frac{1}{N} \sum_{i=1}^N \rho_i \quad (11)$$

wherein the sum runs over the number N of phones in the utterance. This acoustic score 115 ρ is an example of an acoustic-posterior-probability-based score.

The acoustic-posterior-probability-based score 115 ρ is designed to be potentially less affected by changes in spectral match due to particular speaker characteristics or acoustic channel variations. Changes in acoustic match are likely to affect both numerator and denominator similarly in Equation (9), thereby making the acoustic score 115 more invariant to those changes and more focused on phonetic quality.

In the specific embodiment, the acoustic scorer 503 computes each of the context-independent probability densities $p(y|q)$ shown in Equation (9) using distributions from context-independent hidden Markov phone models. In the numerator of Equation (9), $p(y_t|q_i)$ is computed by evaluating the output distribution of the HMM state to which the frame y_t is aligned in phone type q_i 's HMM. The sum over all phone types in the denominator of Equation (9) is computed using the output distribution of the most likely HMM state (for the frame y_t) within each phone type's context-independent HMM.

In the specific embodiment, the output distribution of each state of within each phone type q 's HMM is a weighted mixture of Gaussian distributions. Good results have been achieved using approximately 100 Gaussian distributions with diagonal covariances (i.e., off-diagonal entries constrained to zero in the covariance matrix). Parameter values within

the Gaussian distributions are established, using standard estimation techniques, from training speech data collected from exemplary speakers.

A first alternate acoustic-scoring embodiment computes a context-dependent posterior probability according to a variation of Equation (9). In this embodiment, Equation (9) is replaced by an approximated equation:

$$P(q_i|y_t, ctx_i) \approx \frac{p(y_t|q_i, ctx_i) P(q_i)}{\sum_{\text{All } q} p(y_t|q) P(q)} \quad (12)$$

wherein ctx_i represents phone i 's context class, i.e., the phone types of phone i 's immediately preceding and following phones, as determined by the segmenter HMM 203.

Equation (12) differs from Equation (9) in that the term $p(y_t|q_i, ctx_i)$ in the numerator is computed from an output distribution of an HMM state to which the frame y_t is aligned in a context-dependent (i.e., tri-phone) HMM phone model. This term is the output, or "emission" probability of frame y_t given phone type q_i in the context ctx_i . The denominator still uses the sum over the context-independent phones as in the specific embodiment.

The posterior score ρ_i is replaced (approximated) by a context dependent score ρ'_i , which is defined as the average of the logarithm of the frame-based phone context-dependent posterior probability over all the frames of the segment:

$$\rho_i \approx \rho'_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log P(q_i|y_t, ctx_i) \quad (13)$$

wherein d_i is the duration in frames of the phone i .

The computation may be further simplified; expanding Equation (13) using Equation (12) produces:

$$\rho'_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log [p(y_t|q_i, ctx_i) P(q_i)] - \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log \left[\sum_{\text{All } q} p(y_t|q) P(q) \right] \quad (14)$$

The first term in Equation (14) can be approximated by the log probability per frame along the maximum likelihood path 313

obtained from the the HMM recognizer 203 used for segmentation:

$$\rho_i \approx \rho'_i \approx \log_prob_per_frame_i - \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log \left[\sum_{All q} p(y_t|q) P(q) \right] \quad (15)$$

The context-dependent model used to compute the numerator of Equation (12) is a more precise model than the context-independent one as it captures the realization of the given phone type in the specific phonetic context of the surrounding phones as they occur in the test sentence. Furthermore, the context-dependent score can be faster to compute than the context-independent score, especially if the approximate methods of computation are used. This is true because the many of the context-dependent score's components already exist from operation of the HMM recognizer 203 used for segmentation.

The score for a phone and a sentence are computed similarly as in the specific embodiment, except that in Equation (10), the context-dependent posterior produced by Equation (12) should be substituted for the context-independent posterior produced by Equation (9).

A second alternate acoustic-scoring embodiment is similar to the specific embodiment but the acoustic scorer 503 computes the denominator of Equation (9) by summing over only a subset of the context-independent phones, this reduces computation and allows a similar normalization effect on the acoustic scores, with little reduction in the usefulness of the acoustic score. The phones used are selected to cover most of the acoustic space (i.e., very dissimilar sounds are chosen).

In a third alternate acoustic-scoring embodiment, the acoustic scorer 503 generates the frame-based posterior probabilities $p(q_i|y_i)$ directly by using a multi-layer perceptron (MLP). The multi-layer perceptron is trained using forced (i.e., known-script-constrained) alignments on exemplary training data. The training procedure is a standard backpropagation supervised training scheme.

During training, a current frame--and optionally its surrounding acoustic context frames--is presented to the inputs of the MLP along with the desired output. The desired output for any frame is 1-of-N targets (target 1 is set to the output corresponding to the correct phone type and targets 0 are used for the other outputs). Using a relative entropy or minimum square error training criterion, the outputs are known to converge to the frame-based posterior probabilities $p(q_i|Y_i)$.

MLP's are well known in the art, and are described for example in Nelson Morgan and Herve Bourlard, "Continuous Speech Recognition: An introduction to the Hybrid HMM-Connectionist Approach," IEEE Signal Processing Magazine, Vol. 12, No. 3, May '95, pp. 25-42, which is herein incorporated by reference.

The score for a phone and a sentence are computed similarly as in the specific embodiment, except that in Equation (10), the MLP-based posterior is used instead of the HMM-derived posterior.

In a fourth alternate acoustic-scoring embodiment, the acoustic scorer 503 also generates an acoustic-posterior-probability-based score. However, rather than generating frame-based posterior probabilities according to Equation (9), the acoustic scorer 503 generates phone-based posterior probabilities directly. In this embodiment, the acoustic scorer 503 includes an HMM engine. The acoustic scorer 503 operates the HMM engine to generate an alignment for the frames Y_i of the student speech sample corresponding to a phone i with every phone type q 's hidden Markov phone model using the Viterbi algorithm. The acoustic scorer 503 computes an acoustic log-likelihood, $\log p(Y_i|q)$, of the speech Y_i for each alignment to a phone type q 's HMM using the standard HMM backtracing technique known in the art of speech recognition. Using these log-likelihoods, the acoustic scorer 503 computes a posterior log-probability score for a phone i according to:

$$\rho_i = \log p(q_i | Y_i) = \log \frac{p(Y_i | q_i) P(q_i)}{\sum_{\text{All } q} p(Y_i | q) P(q)} \quad (16)$$

The acoustic scorer 503 computes the acoustic score 115 ρ for an entire utterance as the average of the phone posterior score ρ_i of each phone i in the utterance, according to Equation (11).

IV.B. PHONE LOG-LIKELIHOOD SCORES

In an alternate acoustic-scoring embodiment, the acoustic scorer 503 uses HMM log-likelihoods to derive a likelihood-based pronunciation score 115 L. The underlying assumption is that the logarithm of the likelihood of the speech data, computed by the Viterbi algorithm, using the HMMs obtained from exemplary speakers is a good measure of the similarity (or match) between exemplary speech and the student's speech. The acoustic scorer 503 computes for each phone a normalized log-likelihood l'_i :

$$l'_i = l_i / d_i \quad (17)$$

wherein l_i is the log-likelihood corresponding to phone i and d_i is its duration in number of frames. The normalization by the phone's duration is to give short-duration phones a boost in their effect on the log likelihood score, which would be dominated by longer phones, otherwise.

The acoustic scorer 503 computes the likelihood-based score 115 L for a whole utterance as the average of the individual normalized log-likelihood scores l'_i for each phone i :

$$L = \frac{1}{N} \sum_{i=1}^N l'_i \quad (18)$$

wherein the sum runs over the number N of phones in the utterance.

V. COMBINATION OF SCORES AND MAPPING TO A HUMAN GRADE

FIG. 6 is a block diagram of a system that combines different types of pronunciation scores according to an embodiment of the invention. By combining scores, an improvement in evaluation performance is achieved, overall, as compared to using each score by itself.

In FIG. 6, multiple pronunciation scores 115 are computed for acoustic features 111 of a single utterance. These scores include a phone duration score 115, a syllabic duration score, 115, and an acoustic-posterior-probability-based score 115, which have been described, separately. The scores are shown as being generated by three separate scorers 113. In actual implementation, the three separate scorers 113 would likely share many common components, such as an acoustic segmenter 203 (of FIGS. 2 and 5).

A scores-to-grade mapper 119 accepts the different scores 115 and applies a mapping function 603 to the scores 115 to derive a single grade 121.

FIG. 7 is a block diagram of a system 701 for creating FIG. 6's mapping function 603 between one or more types of machine scores into a pronunciation grade as would be produced by a human listener. In FIG. 7, machine scores 703 are generated for utterances in a development set of training speech data. Human-generated scores 705 are also collected for the utterances in the development set. The development set is assembled so as to include speech from speakers of varying proficiency levels.

A mapping analyzer 707 processes the machine scores 703 and the corresponding human grades 705 to generate a scores-to-grade mapping 603.

In one embodiment of the invention, the mapping analyzer 707 uses linear regression to linearly combine two or more machine scores (m_1, \dots, m_n) for each utterance, plus a bias term, to approximate the corresponding human score h :

$$h' = \lambda_0 + \lambda_1 m_1 + \dots + \lambda_n m_n \quad (19)$$

The linear coefficients λ_i and bias term λ_0 are optimized to minimize the mean square between the predicted and the actual human scores over the utterances of the development set.

In another embodiment of the invention, the mapping analyzer 707 uses nonlinear regression. The machine scores 703 to be combined are the input to a neural network 603 that implements the mapping between the multiple machine scores 703 and the corresponding human scores 705. The mapping analyzer establishes the parameters within the neural network 603 using the actual human scores 705 as targets. The network has a single linear output unit and 16 sigmoidal hidden units. The mapping analyzer trains the neural network using the standard backpropagation technique, using cross-validation on about 15 percent of the training data. The training is stopped when performance degrades on the cross-validation set.

In another embodiment of the present invention, the mapping analyzer 707 computes a mapping 603 that defines the predicted human score h' as the conditional expected value of the actual human score h given the measured machine scores m_1, \dots, m_n :

$$h' = E[h|m_1, \dots, m_n] \quad (20)$$

To compute the expectation the conditional probability $P(h|M_1, \dots, M_n)$ is needed. The mapping analyzer 707 computes this conditional probability as:

$$P(h|m_1, \dots, m_n) = \frac{P(m_1, \dots, m_n|h) P(h)}{\sum_{j=1}^G P(m_1, \dots, m_n|h_j) P(h_j)} \quad (21)$$

wherein the sum in the denominator is over all G possible grades and $P(h)$ is the prior probability of the grade h and the conditional distribution is modeled approximately by a discrete distribution based on scalar or vector quantization of the machine scores. The number of bins to use in the quantization is determined by the amount of available training data. The more available data, the more bins may be used.

In yet another embodiment of the invention, the mapping analyzer 707 uses a decision tree or, alternatively, a class probability tree.

The machine scores to be combined are the input to a tree that implements the mapping between the machine scores 703 and the corresponding human scores 705. The mapping analyzer establishes the parameters within the decision tree (or alternative the class probability tree) using the actual human scores 705 as target classes according to algorithms, known in the art, for constructing decision trees. A discrete set of human targets are defined as classes used by the decision or class probability tree into which classify the input machine scores.

VI. LANGUAGE INSTRUCTION IN A CLIENT-SERVER ENVIRONMENT

FIG. 8 is a block diagram of a distributed system 801 for language instruction that evaluates pronunciation quality. In FIG. 8, a remote client processor 803 runs a client process. The client process executes software instructions that presents a prompt to a student 105. In response, the student 105 speaks into a microphone 805. As will be further discussed, the system 801 contains a pronunciation evaluator (101, as shown in FIG. 1 only). The microphone 805 forms at least a part of the pronunciation evaluator's speech input device (103, as shown in FIG. 1 only).

In one embodiment of FIG. 8, the client process uses a computer display 807 to provide the prompts. One type of prompt is a displayed script to be read by the student 105. The client process exceeds previous pronunciation evaluation systems in that it can (and does) use scripts containing words for which there may be no training data or incomplete training data, as described above. These scripts include scripts generated dynamically during execution by the system 801. Another novel way by which the client process can (and does) elicit the verbal utterances is to ask open-ended questions to which the student 105 answers spontaneously, without reading

from any script, as described above. Thus, the system 801 according to the present invention permits a virtually inexhaustible, immediately-usable supply of unique word sequences for pronunciation evaluation.

In another embodiment, the display 807 is replaced or supplemented by a speaker 809 that provides audio prompts, such as scripts and questions.

A local server processor 811 runs a server process that controls the language instruction lesson being executed on the client processor 803 via a network 813, such as a local area network, the Internet, etc. In one embodiment, the server process controls the lesson by dynamically sending control information that contains or specifies individual prompts, such as scripts and questions, shortly before the prompts are to be provided to the student 105. In another embodiment, the server process controls the lesson more loosely by downloading control information which includes software (e.g., JAVA-language software) for individual lessons to the client processor 803's local storage 815, which includes RAM, or hard disk, etc. The client processor 803 thereafter runs the lesson software with less direct supervision from the server processor 811.

In some embodiments of the invention, the server processor 811 contains the final stages of the pronunciation evaluator which generate the evaluation grade for student pronunciation. In one such embodiment, the microphone 805 is coupled 817 to convey speech to the client processor 803. The client process relays student speech samples across the network 813 to an audio receiver process operating in conjunction with the server process to request pronunciation evaluation. The audio receiver process runs on the server processor 811.

In other such embodiments, the microphone 805 is coupled to relay student speech samples to the server process across a separate channel 819 which is not under the direct control of the client process. The separate channel 819 in one of these embodiments is a *physically* separate channel, such as a telephone channel. The separate channel 819 in

another of these embodiments is a virtual channel that appears to the server process to be a separate channel, even though it is implemented using physical lines also shared by the client-to-server connection. For example, the virtual channel may be implemented using the audio virtual channel of a Digital Simultaneous Voice and Data (DSVD) modem, whose data virtual channel handles the client-to-server communications.

In other embodiments, the pronunciation evaluator (of FIG. 1) is not implemented on the server processor 811. Instead, the evaluator is implemented on either the client processor 803 or elsewhere. Therefore, pronunciation evaluation is controlled by the client process without need for sending speech samples to the server process. In these embodiments, the server processor 811's computation resources are conserved because it needs only control the lesson. In this way, the server processor 811 becomes capable of controlling a greater number of lessons simultaneously in a multi-tasking manner.

As described, the client process and the server process run on separate processors 803 and 811 which are coupled via a network 813. In general, though, the client process and the server process may run on a single processor in a multi-tasking manner.

The invention has now been explained with reference to specific embodiments. Other embodiments will be apparent to those of ordinary skill in the art in view of the foregoing description. For example, preselected scripts may be delivered to a user via off-line means such as a written guidebook, as a newspaper advertisement or in other visual or auditory forms. It is therefore not intended that this invention be limited, except as indicated by the appended claims.

WHAT IS CLAIMED IS:

- 1 1. In an automatic speech processing system, a
2 method for assessing pronunciation of a student speech sample
3 using a computerized acoustic segmentation system, the method
4 comprising:
5 accepting said student speech sample which comprises
6 a sequence of words spoken by a student speaker;
7 operating said computerized acoustic segmentation
8 system to define sample acoustic units within said student
9 speech sample based on speech acoustic models within said
10 segmentation system, said speech acoustic models being
11 established using training speech data from at least one
12 speaker, said training speech data not necessarily including
13 said sequence of spoken words;
14 measuring duration of said sample acoustic units;
15 and
16 comparing said durations of sample acoustic units to
17 a model of exemplary acoustic unit duration to compute a
18 duration score indicative of similarity between said sample
19 acoustic unit durations and exemplary acoustic unit durations.
- 1 2. The method according to claim 1 wherein said
2 exemplary acoustic unit duration model is established using
3 duration-training speech data from at least one exemplary
4 speaker, said duration-training data not necessarily including
5 said sequence of spoken words.
- 1 3. The method according to claim 1 wherein each
2 acoustic unit is shorter in duration than a longest word in
3 the language of said spoken words.
- 1 4. The method according to claim 1 further
2 comprising:
3 mapping said duration score to a grade; and
4 presenting said grade to a student.

1 5. The method according to claim 4 wherein the
2 step of mapping said duration score to a grade comprises:
3 collecting a set of training speech samples from a
4 plurality of language students of various proficiency levels;
5 computing training duration scores for each of said
6 training speech samples;
7 collecting at least one human evaluation grade from
8 a human grader for each of said training speech samples; and
9 adjusting coefficients used in mapping by minimizing
10 an error measurement between said human evaluation grades and
11 said training duration scores.

1 6. The method according to claim 4 wherein the
2 step of mapping comprises using a mapping function obtained by
3 linear or non-linear regression from training duration scores,
4 alone or in combination with other machine scores, and
5 corresponding human evaluation grades, all of said scores and
6 grades being collected over a representative training data
7 base of student speech.

1 7. The method according to claim 6 wherein said
2 mapping function is obtained by non-linear regression
3 implemented with a neural net which allows arbitrary mappings
4 from machine scores to human expert grades.

1 8. The method according to claim 4 wherein the
2 step of mapping comprises using a decision tree or class
3 probability tree whose parameters were established using
4 training duration scores.

1 9. The method according to claim 1 wherein the
2 step of operating said acoustic segmentation system comprises
3 the steps of:
4 computing a path through trained hidden Markov
5 models (HMMs) from among said speech acoustic models, said
6 path being an allowable path through the HMMs that has maximum
7 likelihood of generating an observed acoustic features
8 sequence from said student speech sample; and

9 determining from said path at least one boundary or
10 duration of one acoustic unit.

1 10. The method according to claim 9 wherein:
2 said spoken sequence of words is spoken according to
3 a known script; and
4 the path computing step comprises using said script
5 in defining allowability of any path through the HMMs.

1 11. The method according to claim 9 wherein said
2 spoken sequence of words is unknown, and the path computing
3 step comprises operating a computerized speech recognition
4 system that determines said spoken sequence of words.

1 12. The method according to claim 9 wherein:
2 said sample acoustic units are syllables; and
3 the step of determining at least one acoustic unit
4 boundary or duration comprises the steps of:
5 extracting boundaries or durations of at least
6 two phones from said path; and
7 combining portions of at least two phones to
8 obtain a boundary or duration of a syllable acoustic
9 unit.

1 13. The method according to claim 12 wherein the
2 step of combining portions of at least two phones comprises
3 measuring the time difference between centers of vowel phones
4 from among said phones to obtain a duration of a syllable
5 acoustic unit.

1 14. The method according to claim 1 wherein said
2 sample acoustic units are phones.

1 15. The method according to claim 1 wherein said
2 sample acoustic units are syllables.

1 16. The method according to claim 1 wherein:

2 said exemplary acoustic unit duration distribution
3 model is a model of speaker-normalized acoustic unit
4 durations, and the duration measuring step comprises the steps
5 of:

6 analyzing said student speech sample to
7 determine a student speaker normalization factor; and
8 employing said student speaker normalization
9 factor to measure speaker-normalized durations as said
10 measured sample acoustic unit durations, whereby the
11 comparing step compares said speaker-normalized sample
12 acoustic unit durations to said exemplary speaker-
13 normalized acoustic unit duration distribution model.

1 17. The method according to claim 16 wherein said
2 student speaker normalization factor is rate of speech.

1 18. The method according to claim 1 wherein the
2 step of operating said segmentation system excludes acoustic
3 units in context with silence from analysis.

1 19. The method according to claim 1 wherein the
2 step of operating said segmentation system comprises operating
3 a speech recognition system as said acoustic segmentation
4 system.

* 1 20. A system for assessing pronunciation of a
2 student speech sample, said student speech sample comprising a
3 sequence of words spoken by a student speaker, the system
4 comprising:

5 speech acoustic models established using training
6 speech data from at least one speaker, said training speech
7 data not necessarily including said sequence of spoken words;
8 a computerized acoustic segmentation system
9 configured to identify acoustic units within said student
10 speech sample based on said speech acoustic models;
11 a duration extractor configured to measure duration
12 of said sample acoustic units;
13 a model of exemplary acoustic unit duration; and

14 a duration scorer configured to compare said sample
15 acoustic unit durations to said model of exemplary acoustic
16 unit duration and compute a duration score indicative of
17 similarity between said sample acoustic unit durations and
18 acoustic unit durations in exemplary speech.

1 21. In an automatic speech processing system, a
2 method for grading the pronunciation of a student speech
3 sample, the method comprising:
4 accepting said student speech sample which comprises
5 a sequence of words spoken by a student speaker;
6 operating a set of trained speech models to compute
7 at least one posterior probability from said speech sample,
8 each of said posterior probabilities being a probability that
9 a particular portion of said student speech sample corresponds
10 to a particular known model given said particular portion of
11 said speech sample; and
12 computing an evaluation score, herein referred to as
13 the posterior-based evaluation score, of pronunciation quality
14 for said student speech sample from said posterior
15 probabilities.

1 22. The method according to claim 21 wherein each
2 of said posterior probabilities is derived from a model
3 likelihood by dividing the likelihood that said particular
4 known model generated said particular portion of said student
5 speech sample by the summation of the likelihoods that
6 individual models generated said particular portion of said
7 speech sample.

1 23. The method according to claim 21 wherein:
2 said particular known model is a context-dependent
3 model; and
4 said individual models are context-dependent or
5 context-independent models.

1 24. The method according to claim 21 wherein:

2 said trained speech models comprise a set of phone
3 models;
4 said student speech sample comprises phones; and
5 the step of operating said speech models comprises
6 computing a frame-based posterior probability for each frame
7 y_i within a phone i of a phone type q_i :

$$P(q_i|y_t) = \frac{P(y_t|q_i, \dots) P(q_i)}{\sum_{\text{All } q} P(y_t|q, \dots) P(q)}$$

8

9 wherein:

10 $p(y_t|q_i, \dots)$ is the probability of the frame y_t
11 according to a model corresponding to phone type q_i ;
12 the sum over q runs over all phone types; and
13 $P(q_i)$ represents the prior probability of the
14 phone type q_i .

1 25. The method according to claim 24 wherein the
2 step of computing a frame-based posterior probability uses
3 context-dependent models corresponding to each phone type q_i
4 in the numerator, whereby said $p(y_t|q_i, \dots)$ is a context-
5 dependent likelihood $p(y_t|q_i, \text{ctx}_i)$, wherein ctx_i represents
6 context.

1 26. The method according to claim 24 wherein the
2 step of computing said posterior-based evaluation score for
3 said student speech sample comprises computing for a phone i
4 an average of the logarithm of the frame-based posterior
5 probabilities of all frames within said phone i , said average
6 herein referred to as a phone score ρ_i , which is expressible
7 as:

$$\rho_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log P(q_i|y_t)$$

8 wherein the sum runs over all d_i frames of said phone i .

1 27. The method according to claim 26 wherein said posterior-
2 based evaluation score for said student speech sample is
3 defined as an average of the individual phone scores ρ_i for
4 each phone i within said student speech sample:

$$\rho = \frac{1}{N} \sum_{i=1}^N \rho_i$$

5 wherein the sum runs over the number of phones in said student
6 speech sample.

1 28. The method according to claim 24 wherein the
2 model corresponding to each phone type is a gaussian mixture
3 phone model.

1 29. The method according to claim 24 wherein the
2 model corresponding to each phone type is a context-
3 independent phone model.

1 30. The method according to claim 24 wherein the
2 model corresponding to each phone type is a hidden markov
3 model.

1 31. The method according to claim 22 wherein said
2 particular portion of said speech sample is a phone.

1 32. The method according to claim 21 further
2 comprising:
3 mapping said posterior-based evaluation score to a
4 grade as would be assigned by human listener; and
5 presenting said grade to said student speaker.

1 33. The method according to claim 32 wherein said
2 step of mapping said posterior-based evaluation score to a
3 grade comprises:

4 collecting a set of training speech samples
5 from a plurality of language students of various
6 proficiency levels;

7 collecting a set of human evaluation grades for
 8 each of said training samples from human expert listeners
 9 listening to said samples; and
 10 adjusting coefficients used in mapping by
 11 minimizing the squared-error between the human expert
 12 grades and said evaluation score.

1 34. The method according to claim 21 wherein said
 2 student speech sample comprises an acoustic features sequence,
 3 the method further comprising the steps of:

4 computing a path through a set of trained hidden
 5 Markov models (HMMs) from among said speech acoustic models,
 6 said path being an allowable path through the HMMs that has
 7 maximum likelihood of generating said acoustic features
 8 sequence; and

9 identifying transitions between phones within said
 10 path, thereby defining phones.

1 35. The method according to claim 34 wherein the
 2 path computing step is performed using the Viterbi search
 3 technique.

1 36. The method according to claim 34 wherein said
 2 spoken sequence of words is unknown, and the path computing
 3 step is performed using a computerized speech recognition
 4 system that determines said spoken sequence of words.

1 37. The method according to claim 21 wherein
 2 segments in context with silence are excluded from said
 3 student speech sample and from training data used to train
 4 said speech models.

* 1 38. A system for assessing pronunciation of a
 2 student speech sample, said student speech sample comprising a
 3 sequence of words spoken by a student speaker, the system
 4 comprising:
 5 trained speech acoustic models of exemplary speech;
 6 and

7 an acoustic scorer configured to compute at least
8 one posterior probability from said speech sample using said
9 trained speech models, said acoustic scorer also configured to
10 compute an evaluation score of pronunciation quality for said
11 student sample from said posterior probabilities, each of said
12 posterior probabilities being a probability that a particular
13 portion of said student speech sample corresponds to a
14 particular known model given said particular portion of said
15 speech sample.

1 39. A system for pronunciation training in a
2 client/server environment wherein there exists a client
3 process for presenting prompts to a student and for accepting
4 student speech elicited by said prompts, the system
5 comprising:

6 a server process for sending control information to
7 said client process to specify a prompt to be presented to
8 said student and for receiving a speech sample derived from
9 said student speech elicited by said presented prompt; and
10 a pronunciation evaluator invocable by said server
11 process for analyzing said student speech sample.

1 40. The system according to claim 39 wherein:
2 said pronunciation evaluator is established using
3 training speech data; and
4 said server process is adapted to specify a prompt
5 for eliciting a sequence of words not necessarily found in
6 said training speech data as said student speech sample.

1 41. The system according to claim 39 wherein said
2 server process receives said speech sample over a speech
3 channel that is separate from a communication channel through
4 which said server process and said client process communicate.

1 42. The system according to claim 39 wherein said
2 client process and said server process are located on two
3 separate computer processors and communicate via a network.

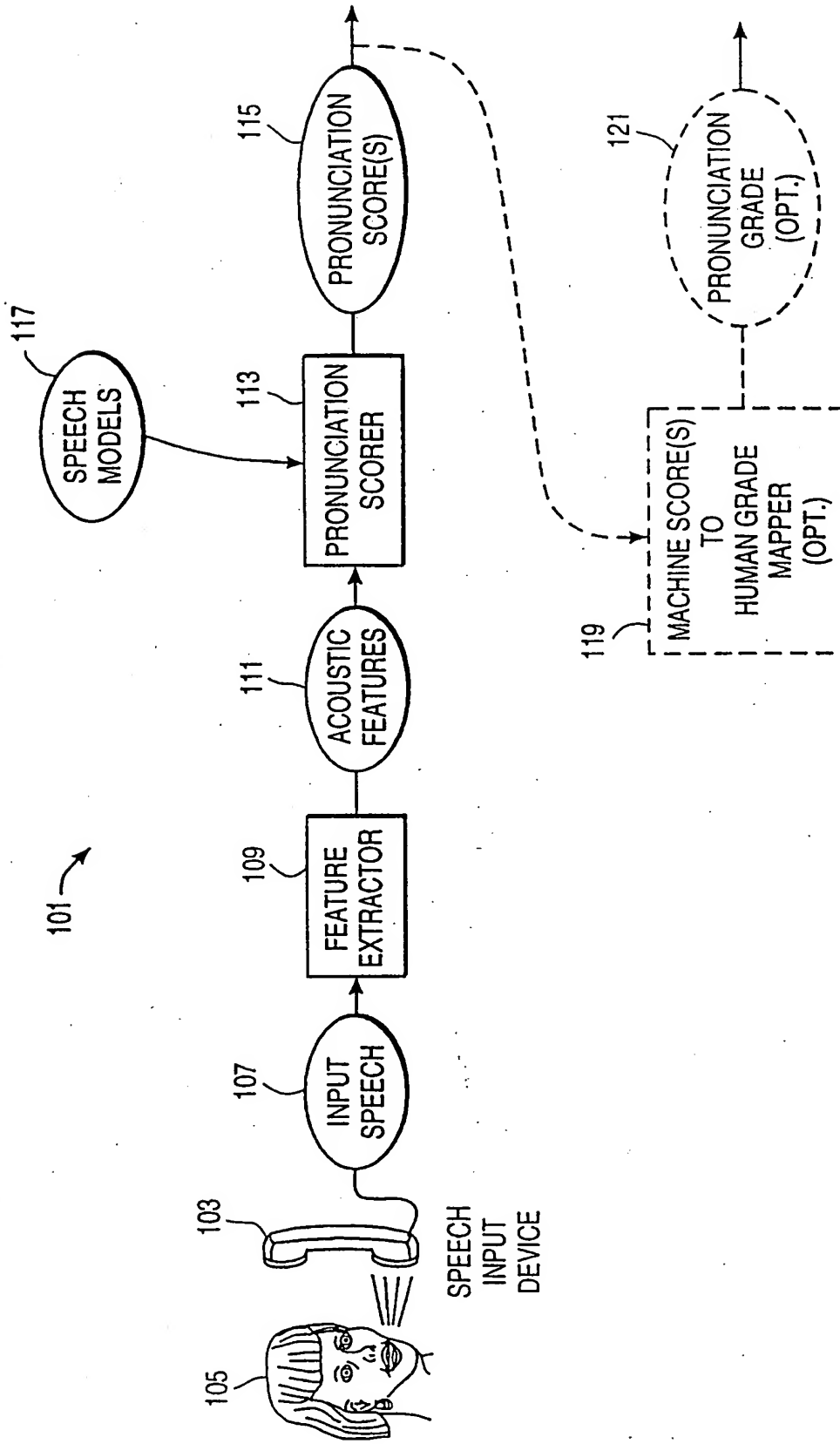


FIG. 1

2/8

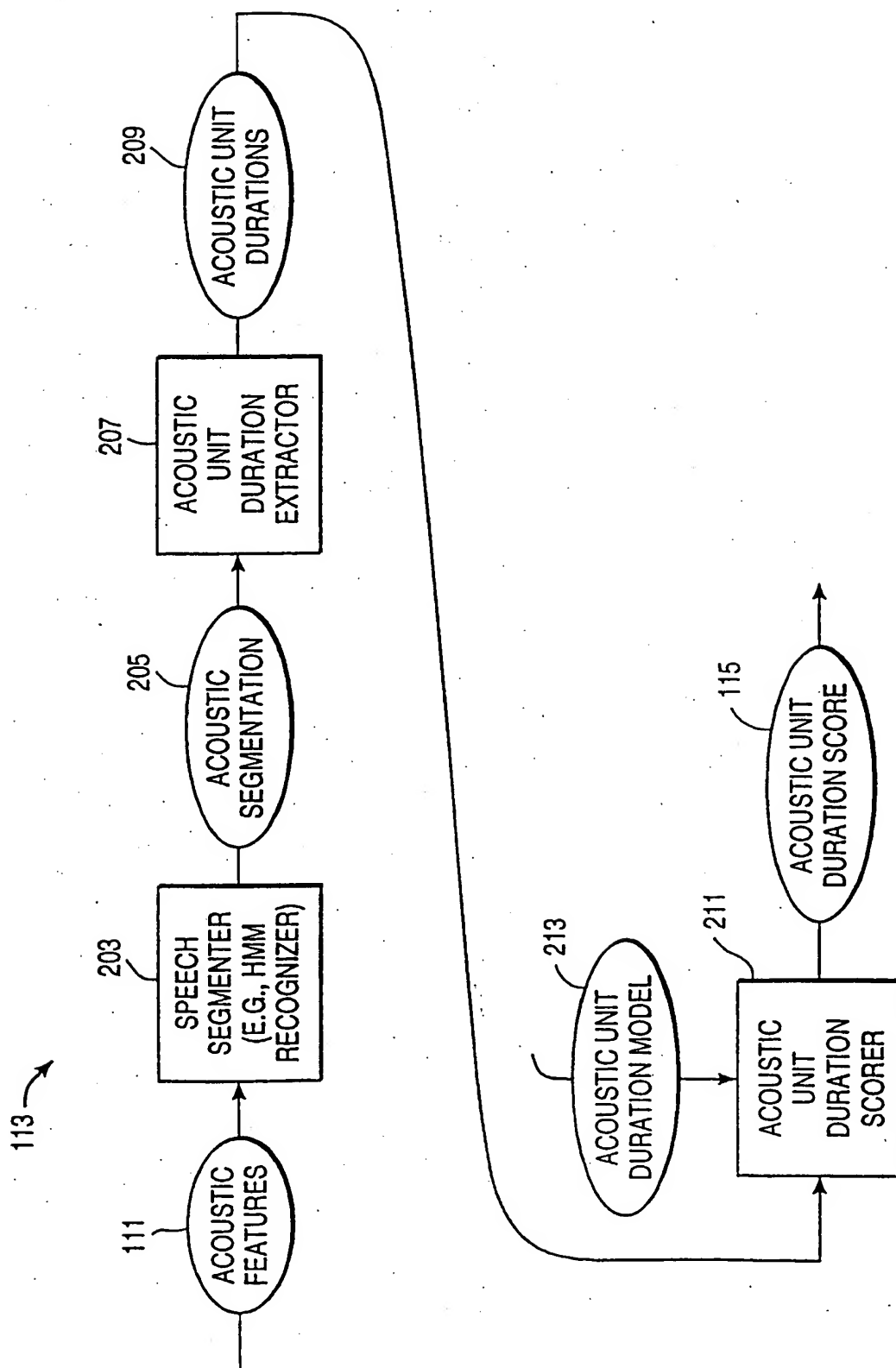


FIG. 2

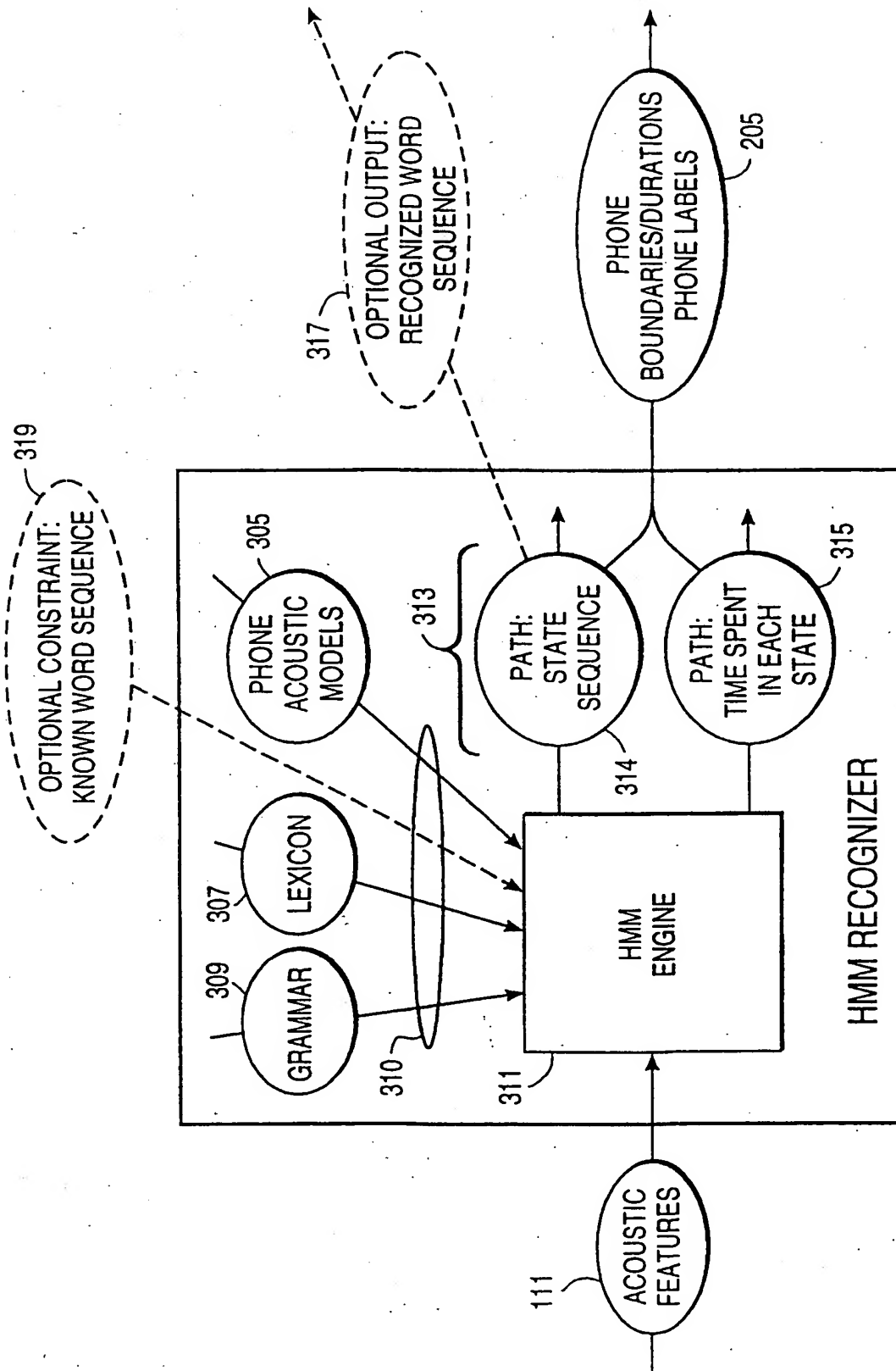


FIG. 3

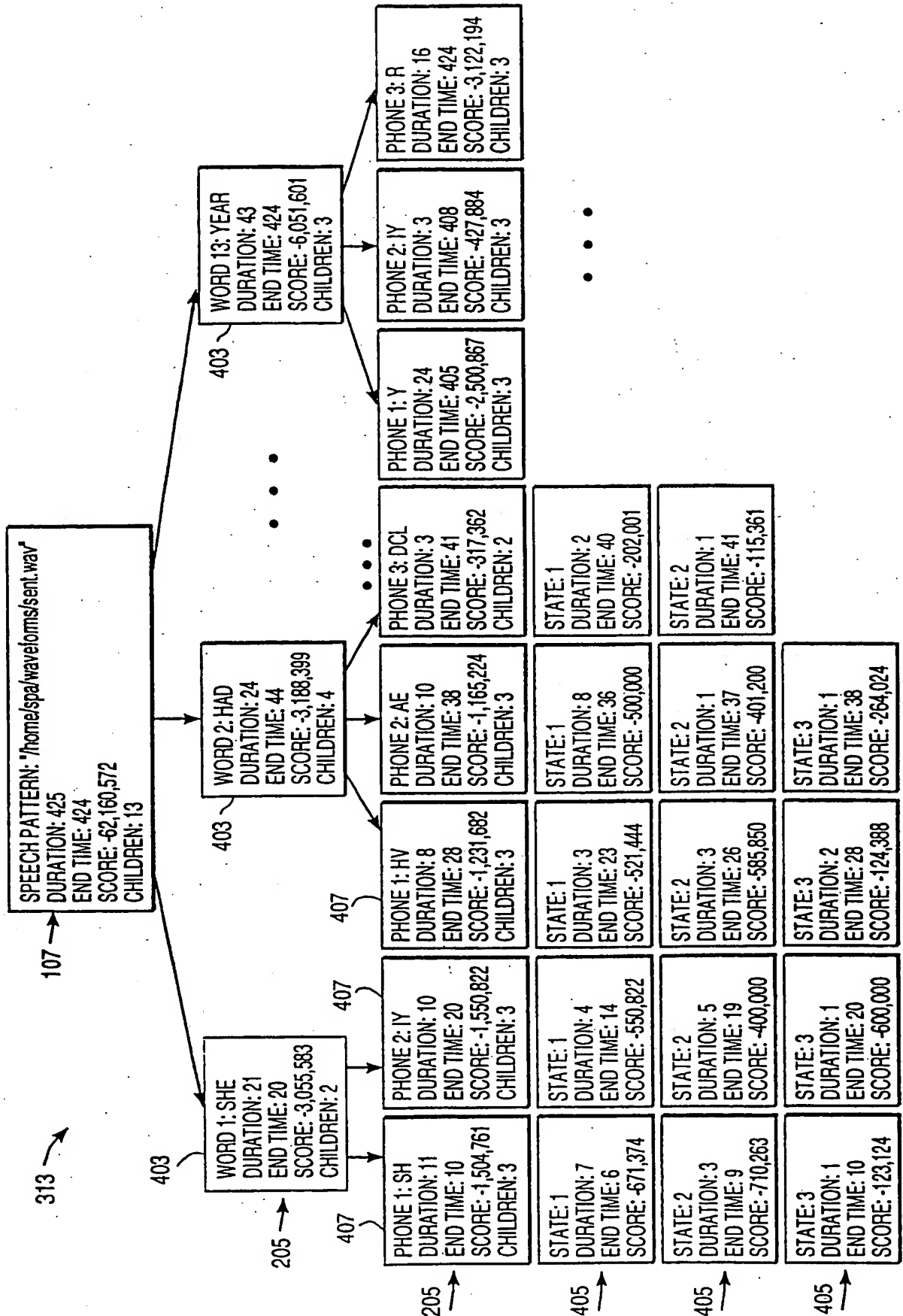


FIG. 4

113 →

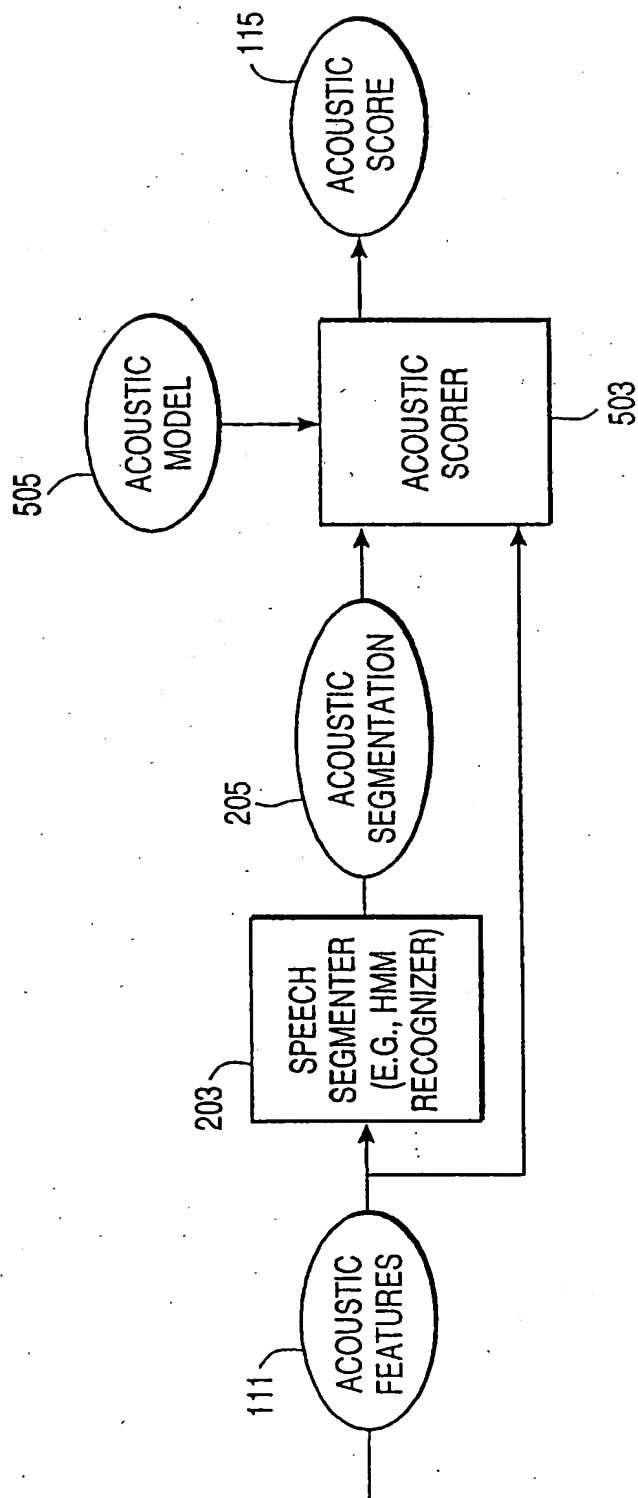


FIG. 5

6 / 8

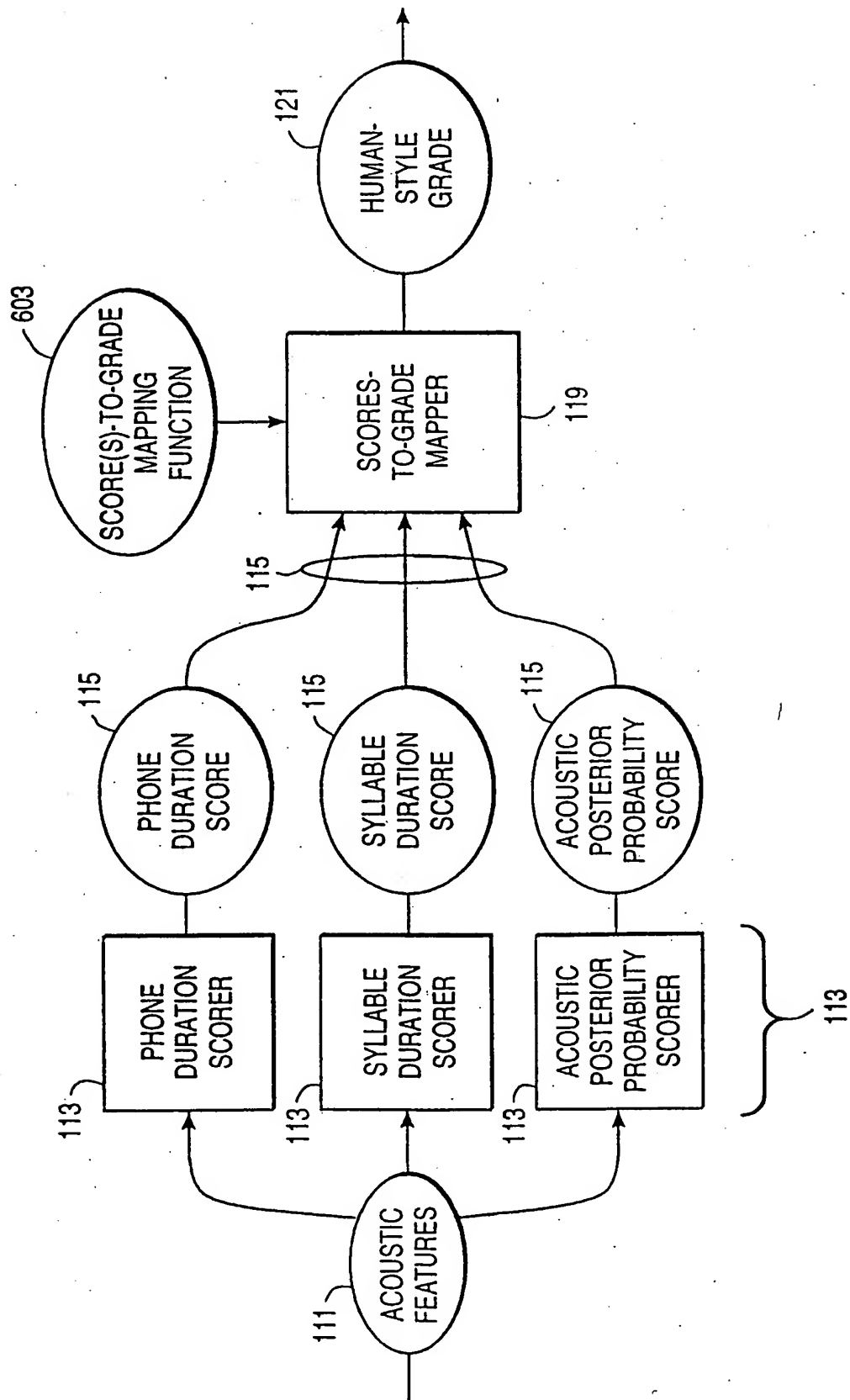


FIG. 6

8 / 8

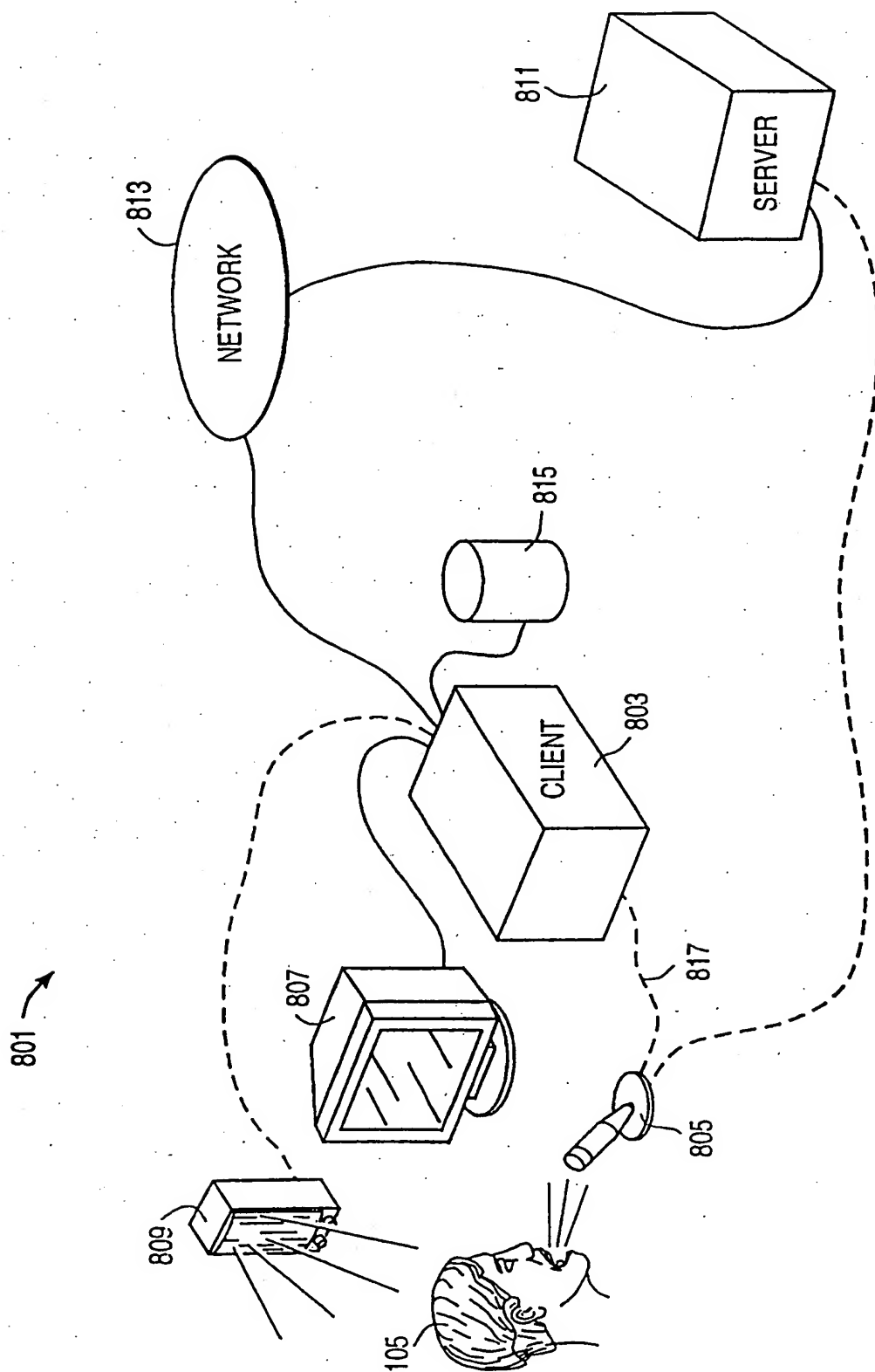


FIG. 8

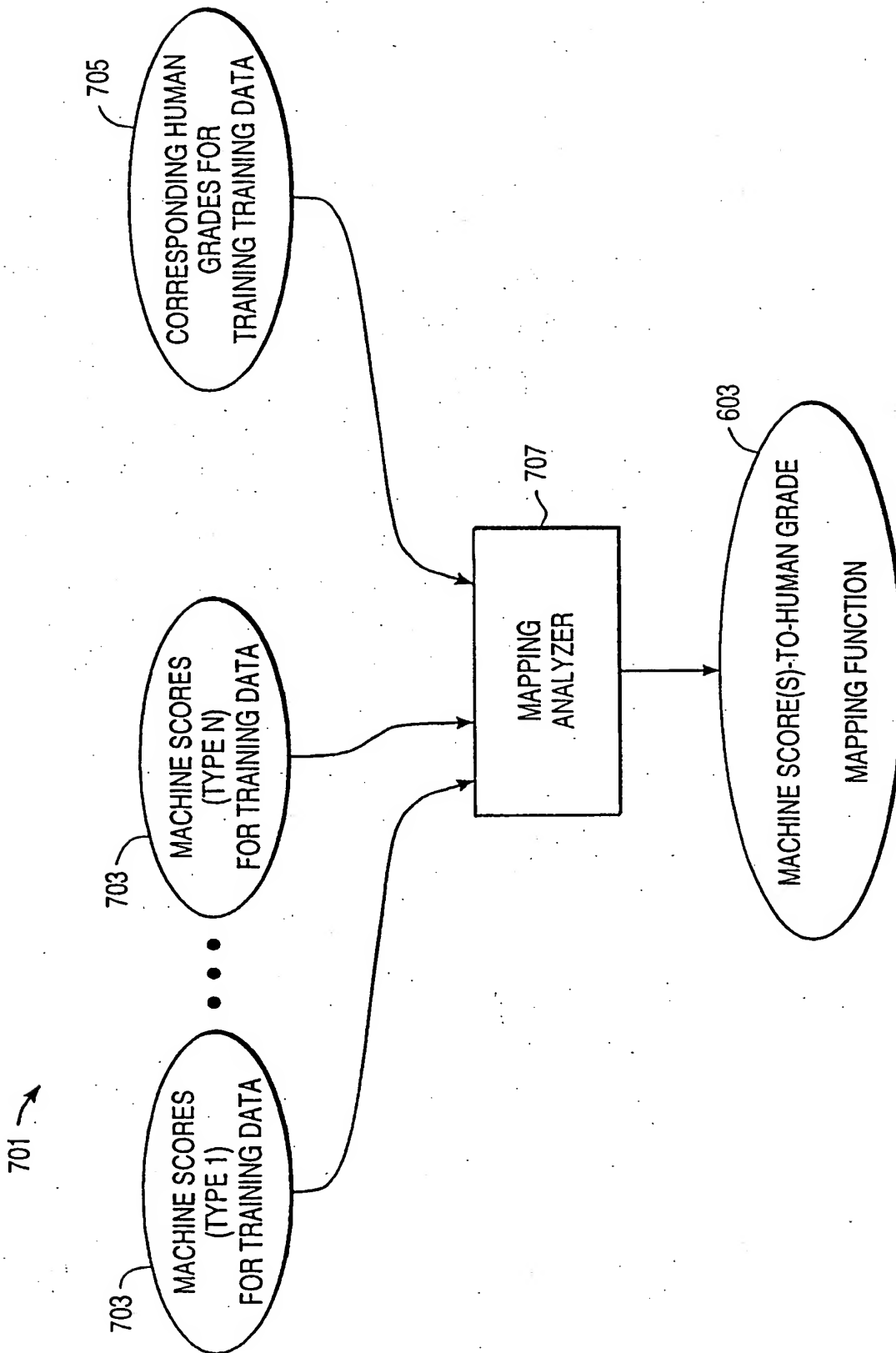


FIG. 7

8 / 8

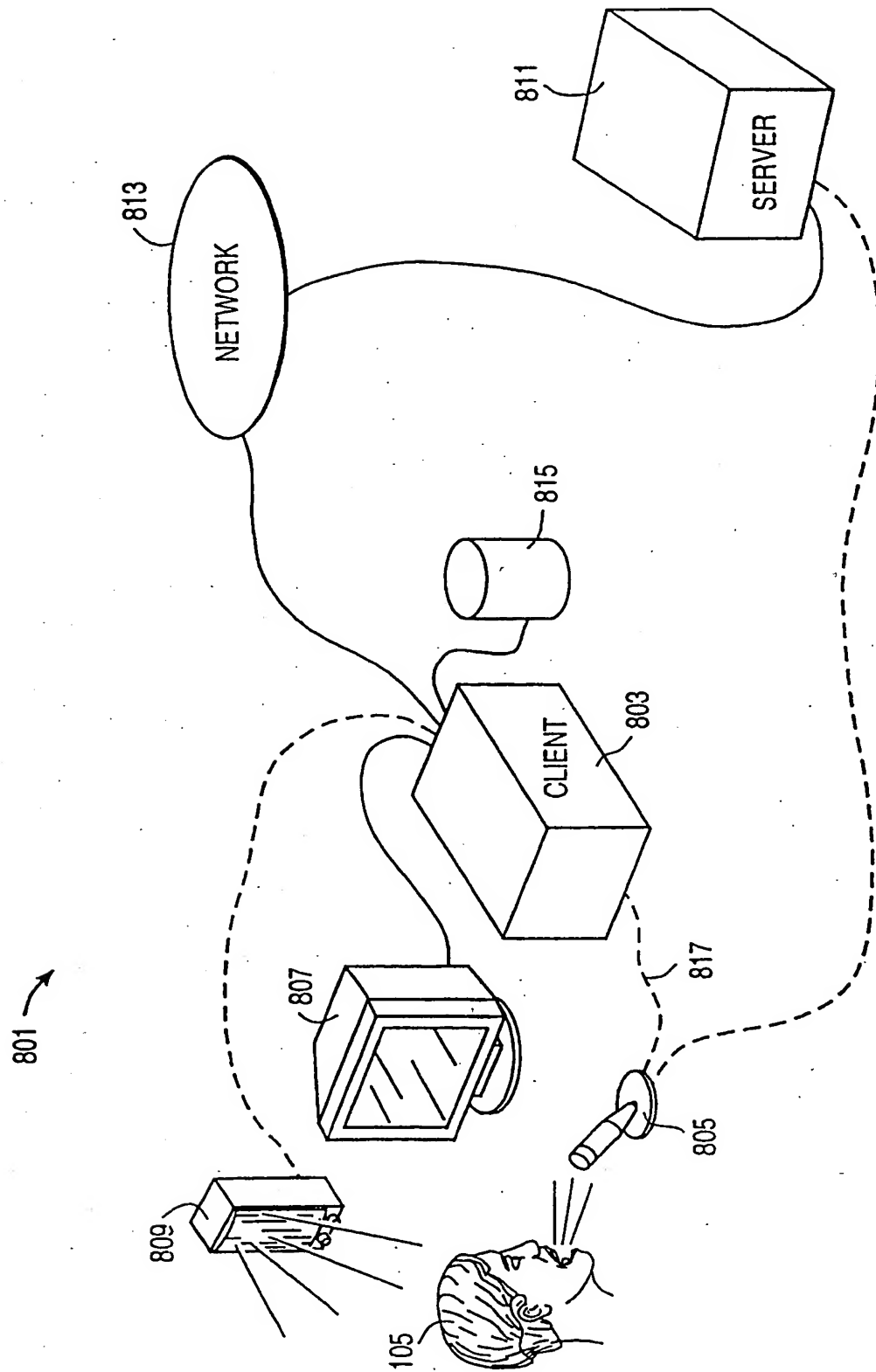


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US97/17888

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G10L 5/04, 9/00; G09B 5/00.

US CL : 704/211, 254, 256, 270; 434/185.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 704/211, 254, 256, 270; 434/185.

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS search terms: pronunciation, speech, voice, duration, probability, Markov, time?, map?, and grad?.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X — Y	US 4,969,194 (EZAWA et al) 06 November 1990, see especially columns 2 and 14.	1-3, 14-20 — 4-13
X — Y	US 5,487,671 (SHPIRO et al) 30 January 1996, see especially columns 1-2 and 7.	1-3, 14-20 — 4-13
X — Y	US 5,268,990 (COHEN et al) 07 December 1993, see especially abstract, figure 5, columns 1, 5 and 6.	1-23, 31-38 — 24-30
X	US 4,615,680 (TOMATIS) 07 October 1986, see especially claim 4.	1-3, 14-20

☒ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

* Special categories of cited documents:	* T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
* A* document defining the general state of the art which is not considered to be of particular relevance	* X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
* B* earlier document published on or after the international filing date	* Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
* L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	* A* document member of the same patent family
* O* document referring to an oral disclosure, use, exhibition or other means	
* P* document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 06 DECEMBER 1997	Date of mailing of the international search report 10 FEB 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer <i>B. Knepper</i> DAVID D. KNEPPER Telephone No. (703) 305-9644

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US97/17888

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 4,472,833 (TURRELL et al) 18 September 1984, see abstract and column 3.	1-3
X, E ----- Y, E	US 5,679,001 (RUSSELL et al) 21 October 1997, see especially abstract, columns 6, 7, and 12.	9-13, 21-23, 31-38 ----- 1-8, 14-20
X, P ----- Y, P	US 5,581,655 (COHEN et al) 03 December 1996, see especially abstract and columns 1, 11 and 12.	1-23, 31-38 ----- 24-30
X, P ----- Y, P	US 5,634,086 (RTISCHEV et al) 27 May 1997, see especially abstract and col 4.	9-13, 21-23 and 31-38 ----- 39-42
Y	US 3,881,059 (STEWART) 29 April 1975, see especially figures 1 and 2.	39-42
Y	US 3,453,749 (SNEDEKER) 08 July 1969, see figures 1-3.	39-42
Y	US 3,184,549 (AUERNHEIMER) 18 May 1965, see figure 1.	39-42
Y	US 4,799,261 (LIN et al) 17 January 1989, see abstract and figure 1.	1-20
Y	US 5,455,889 (BAHL et al) 03 October 1995, see abstract and figure 4.	1-38

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/17888

Box I Observations where certain claims were found unsearchable (Continuation of Item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of Item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐
☐

The additional search fees were accompanied by the applicant's protest.

No protest accompanied the payment of additional search fees.